REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
15/04//2014	Quarterly Interim Research Performance Report	July 2014 – September 2014

4. TITLE AND SUBTITLE

C3: The Compositional Construction of Content
A new, more effective and efficient way to marshal inferences from background knowledge that will enable more natural and effective communication with automomous systems

5a. CONTRACT NUMBER

5b. GRANT NUMBER

N00014-13-1-0228

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Dr. David D. McDonald
Prof. James D. Pustejovsky

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Smart Information Flow Technologies, dba SIFT, LLC

8. PERFORMING ORGANIZATION REPORT NUMBER

C3-Q-12-2013

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Office of Naval Research
875 North Randolph Street
Arlington, VA 22203-1995

10. SPONSOR/MONITOR'S ACRONYM(S)

ONR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for Public Release,

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Reports on publications and presentations done during the quarter. Summarizes research plan going forward. Describes new research questions identified during the quarter.

15. SUBJECT TERMS

Deep natural language understanding, efficient inference, pragmatics, background knowledge

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr. David McDonald
U	U	U	SAR	4	19b. TELEPHONE NUMBER *(Include area code)* (781) 718-1964

Quarterly Interim Research Performance Report-
July – September 2014

C3: The Compositional Construction of Context

'A more effective and efficient way to marshal inferences from background knowledge'

N00014-13-1-0228

Dr. David McDonald
Smart Information Flow Technologies, dba SIFT, LLC
dmcdonald@sift.net

with Prof. James Pustejovsky
Brandeis University
jamesp@cs.brandeis.edu

1. What we've done

Much of our effort this quarter was focused on preparations for the ONR annual review at the mid-point of this period. In addition Prof. Pustejovsky gave several extended presentations on events, situations, and habitats, and we began to work out how to apply background knowledge in the problem of inferring missing arguments, the problem that is likely to be the focus of our attention in the next quarter.

Research The notion of a *lexicalized ontology* is central to our approach. The immediate link between the use of a word and the evocation of its counterpart in the ontology lets us deploy their associated inferences efficiently and in a manner tailored to the ongoing situation.

Before this period, we focused on the link between single words and the complex habitats they evoked. The texts that we used at that time lent themselves to a strictly compositional analysis where the semantic contribution of each word could be incorporated into the model of the situation at the moment it was reached in the analysis. However, word-at-a-time operations are not practical and perhaps not even possible in the new corpus of biomedical text that we have started to use for our C3 research (see below). The problem is that much of the biomedical vocabulary involves general concepts (*load*, *activate*) that will only get a specific meaning when we see them in construction with their actual arguments in the text, only then can we invoke or elaborate the appropriate habitat (frame).

Some of the underspecification can be resolved locally. Biologists typically use the same name for a gene and the protein that it expresses, e.g., *Ras*. This is an instance of a logically polysemous type that follows the well-established general pattern of producer and product. The gene is the producer and the protein is the product. Which of these alternatives is intended in a given instance is easily determined from the immediate context because proteins and genes take part in quite different activities. Only genes can mutate, so when we read "'*the most prevalent oncogenic mutations in Ras*" the selectional restriction on the possible values for the predicate **mutate** will select the Ras gene. Alternatively, in a text such as "*GTP hydrolysis on Ras*" the protein is selected because **hydrolysis** is a biological process that only occurs on (portions of) proteins or similar molecules.

Other kinds of underspecification, particularly what we refer to as the problem of *missing arguments*, by their nature cannot be resolved locally. Consider the sentence "*Ras acts as a molecular switch that is activated upon GTP loading*[1] *and deactivated upon hydrolysis*[2] *of GTP to GDP*." The events described by the two verbs marked with superscripts are syntactically correct but logically incomplete. They are both describing an operation involving a so-called small molecule, in phrase 1 it is being added and in phrase 2 it is being removed, but added and removed from what?

The syntactic relationship between the main clause of that sentence and its two *upon* adjuncts is not the sort that carries entities from the main clause into its adjuncts.[1] Instead we need to get the information from the **switches habitat** that is activated by the phrase *acts as a molecular switch*, or in a mature model simply by the reference to the protein Ras.

Presentations and publications We completed the editorial process on our submission to the Advances in Cognitive System journal, so our paper "Representing Inferences and their Lexicalization" is now published and can be downloaded from `http://www.cogsys.org/journal/volume-3/`. Full publication details are given below. We had to remove a considerable amount from our original draft, so we anticipate issuing a technical report where that material will be restored.

Professor Pustejovsky's paper with our grant-supported graduate student Nikhil Krishnaswamy, "Cenerating Simulations of Motion Events from Verbal Descriptions," was delivered in August at the 3d Joint Conference on Lexical and Computational Semantics (*SEM 2014). Publication details below.

On July 7th, Pustejovsky gave a Plenary lecture in Prague at the John's Hopkins Center for Language and Speech Processing (`https://ufal.mff.cuni.cz/JHU-PIRE-workshop-2014`). as part of the Fred Jelinek Memorial Workshop. The title of the talk was "Distinguishing 'possible' from 'probable' meaning shifts: How distributions impact linguistic theory."

> In this talk, I discuss the changing role of data in modeling natural language, as captured in linguistic theories. The generative tradition of introducing data using only "evaluation procedures", rather than "discovery procedures", promoted by Chomsky in the 1950s, is slowly being unraveled by the exploitation of significant language datasets that were unthinkable in the 1960s. Evaluation procedures focus on possible generative devices in language without constraints from actual (probable) occurrences of the constructions. After showing how both procedures are natural to scientific inquiry, I describe the natural tension between data and the theory that aims to model it, with specific reference to the nature of the lexicon and semantic selection. The seeming chaos of organic data inevitably violates our theoretical assumptions. But in the end, it is restrictions apparent in the data that call for postulating structure within a revised theoretical model.

1. Compare that sentence pattern to the so-called control constructions: "*Bob persuated Alice to come with him*," where the the subject of the infinitive complement *to come* is guarenteed to be the same as subject of the upstairs clause.

At the end of August, Pustejovsky gave a well received presentation at the Concept Types and Frames workshop in Dusseldorf (`http://www.sfb991.uni-duesseldorf.de/ctf-2014/`).

> In this talk I examine recent work in cognitive science and linguistics arguing that language interpretation involves the creation of a simulation of the utterance. Some of those developing such a view include Barselou (1999), Feldman and Narayanan (2003), Evans (2008), and Bergen (2013). Experimental evidence from psycholinguistic studies increasingly support such a view, and some linguists are working to accommodate these findings theoretically, e.g., Evans (2008). Here, I will generally agree with this program of research. Still missing from these accounts, however, is a formal or computational characterization of what a simulation is, and how it is constructed. This is important if the theory is to be tested and evaluated against the same linguistic data and phenomena as other linguistic theories. I outline what such a model of simulation generation should look like, and how it compares to formal theories of semantics for natural language.

2. What we're planning to do

Corpus From the perspective of our task — determining how to effectively deploy the background knowledge we all use when we are listening or reading — biomedical texts provide what could be described as a target-rich environment. In their research papers (as opposed to their textbooks), Biologists presume that their readers already have a significicant amount of knowledge about the subject. As a result, they leave it to the reader to make the "obvious" connections because they know that their target readers (other biologists) will infer the values of the missing links. The result is that every sentence and nearly every clause in a biomedical article contains logical gaps like the ones described earlier.

Ontology There are a great many open questions about what precisely has to happen during the semantic interpretation of a knowledge-rich text to effectively marshal the inferences that make it possible to understand it. Given the context-dependency of the verbs that we aluded to, we want to explore the use of partially saturated terms as a possible locus for inference. This would be both the simple inferences that identify the correct meaning of an underspecified general verb such as *load*, which takes on a meaning roughly the equivalent of 'form a molecular bond between' when it is in composition with a protein, as in *GTP loading*. And the broader inferences — the 'bringing to mind' of a large body of background knowledge, mediated by a habitat — whereby an instance of that partially-saturated phrase also evokes the Ras-based molecular switch that the loading action turns on and the downstream effects of its activation.

We use the KRISP knowledge representation[2] as the basis of our work. It includes a first-class representation of partially-saturated individuals, and a scheme for reifying classes of such individals as so-called *derived categories* when there is a need to predicate facts about them beyond their immediate content. For example, a title such as *senior vice-president* usually appears as part of the three place predicate idenfying the person holding the position and company at which they work. But in a text like "*Senior vice presidents at IBM have signing authority up to $300,000.*" there is no mention of a particular person, only of a derived category, where two of the variables in the **position** category are bound to particular individuals and the **person** variable is free.

There has yet to be a satisfactory implementation of partially-saturated individuals or derived categories in KRISP. We intend to address that this quarter given the rich set of examples we can now use to establish their epistiemological characteristics. In particular, we think that a defeasible

2. McDonald, David D. (2000) *Issues in the Representation of Read Texts: The Design of Krisp* in Iwanska & Shapiro (eds.) **Natural Language Processing and Knowledge Representation**, MIT Press, 77–110.

binding of the open variable in the representation of the meaning of, e.g., *GTP loading* as a derived category could provide a natural link to the habitat that it evokes.

Publication References

McDonald, David D. & Pustejovsky, James P. *Representing Inferences and their Lexicalizaton*, Advances in Cognitive Systems, 2014, (3) 143-162.

Pustejovsky, James P. & Nikhil Krishnaswamy *Cenerating Simulations of Motion Events from Verbal Descriptions'*, Proceedings of the Third Joint Conference on Lexical and Computational Semantics, 99-109, Dublin, Ireland, August 23-24 2014.